Biost 517
Applied Biostatistics I

Midterm Examination Key
November 6, 2009

Name: _____ Disc Sect: M W F

Instructions: Please provide concise answers to all questions. The exam is worth a total of 137 points.

Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.

The examination is closed book and closed notes. You may use calculators, but you may not use any special programs written for programmable calculators.

If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor on Monday.

PLEDGE:

On my honor, I have neither given nor received unauthorized aid on this examination:

Signed:_____

Problems 1 – 2 relate to a study of 735 patients undergoing brain imaging (MRI) in order to study associations between central nervous system disease and cardiovascular disease. The following variables are available:

- **ptid=** Subject identification number
- **mridate=** Date of enrollment in the study (MMDDYY format)
- **age=** Age (years)
- **male =** Sex (0=female, 1=male)
- **bmi=** Body mass index (weight / height^2) (kg / m^2)
- **packyrs=** = Participant smoking history in pack years (1 pack year = smoking 1 pack of cigarettes per day for 1 year). A participant who never smoked has 0 pack years.
- **physact=** Physical activity of the participant for the week prior to MRI (measured in 1,000 kcal)
- **genhlth=** General state of health (1= best, 2, 3, 4, or 5= worst)
- **ldl=** Blood levels of low density lipoprotein ("bad cholesterol") (mg/dl).
- **sbp =** Systolic blood pressure (mm Hg)
- **obstime=** Observation time from enrollment until death or until end of study, whichever comes first (years)
- **status=** Survival status at time of last observation time (0=still alive, 1= dead

The following table contains descriptive statistics on the sample.

Variable	N	Mean	SD	Min	25th %ile	Median	75th %ile	Max
ptid	735	368	212.3	1	184	368	552	735
mridate	735	76423	31896	10192	63092	80992	91392	123191
age	735	74.6	5.5	65	71	74	78	99
male	735	0.498	0.500	0	0	0	1	1
bmi	735	26.4	4.32	14.5	23.6	26.0	28.5	46.7
packyrs	734	19.6	27.1	0	0	6.5	33.8	240
physact	735	1.92	2.05	0.00	0.55	1.31	2.52	13.82
genhlth	735	2.59	0.94	1	2	3	3	5
ldl	725	125.8	33.6	11	102	125	147	247
sbp	735	131.1	19.7	78	118	130	142	210
obstime	735	4.94	1.07	0.19	5.03	5.14	5.60	5.91
status	735	0.181	0.385	0	0	0	0	1

1. (3 points each part) For each of the following variables, circle the descriptive statistics that are **NOT** of use to provide a scientifically meaningful description of the sample. Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: In the answers below, I underline and put in boldface type the descriptive statistics that are not of use.

a. Consider **ptid**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

The variable ptid is a nominal variable (coded as a number). *No arithmetic makes sense.*

b. Consider **mridate**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

The variable mridate has to be considered a nominal variable the way it is coded (MMDDYY). *No arithmetic makes sense.*

c. Consider **age**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean *Std Dev* *Minimum* *25th Pctile* *Median* *75th Pctile* *Maximum*

The variable age is a continuous variable. *(It is typically measured discretely, but is continuous.)*

d. Consider **male**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

The variable male is a binary variable. *The mean is the proportion of males, all the others are boring because they only give the same information as the mean, but in a more complicated way.*

e. Consider **bmi**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean *Std Dev* *Minimum* *25th Pctile* *Median* *75th Pctile* *Maximum*

The variable age is a continuous variable.

f. Consider **packyrs**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean *Std Dev* *Minimum* *25th Pctile* *Median* *75th Pctile* *Maximum*

The variable packyrs is a continuous variable. *(It is often measured discretely, but is continuous.)*

g. Consider **physact**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean *Std Dev* *Minimum* *25th Pctile* *Median* *75th Pctile* *Maximum*

The variable physact is a continuous variable.

h. Consider **genhlth**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean **Std Dev** *Minimum* *25th Pctile* *Median* *75th Pctile* *Maximum*

The variable genhlth is an ordered categorical variable. *Because ordering makes sense, we can use the extrema and the quantiles. But there is no way to interpret the mean and standard deviation by themselves.*

i. Consider **ldl**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean *Std Dev* *Minimum* *25th Pctile* *Median* *75th Pctile* *Maximum*

The variable ldl is a continuous variable.

j. Consider **sbp**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean *Std Dev* *Minimum* *25th Pctile* *Median* *75th Pctile* *Maximum*

The variable sbp is a continuous variable.

k. Consider **obstime**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

The variable obstime is a right censored continuous variable. *We would need to use Kaplan-Meier estimates to get descriptive statistics about survival times.*

l. Consider **status**. Circle the descriptive statistics that are NOT useful. Briefly explain why.

Mean **Std Dev** **Minimum** **25th Pctile** **Median** **75th Pctile** **Maximum**

The variable status is a binary variable indicating uncensored observations. *While the mean tells us the proportion that are uncensored, we do not know the timeframe of those measurements without also considering obstime. For useful descriptive statistics about patient survival, we would need to use obstime and status together in a Kaplan-Meier estimate.*

2. (5 points) How would your answers to problem 1 change if we were trying to compare distributions across populations, instead of just describing the sample. Briefly explain your reasons.

Ans: We generally do not find it useful to compare minima or maxima across populations, as they are too dependent on the sample size.

For the ordered categorical variable genhlth, we can consider comparing means across populations, in which case the standard deviation is also of use in deriving the samping distribution.

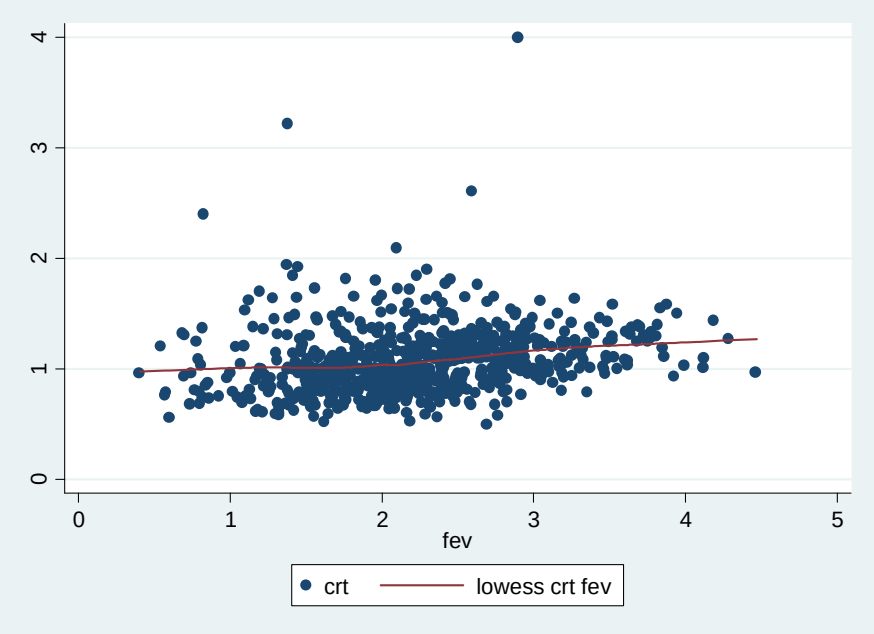
3. (10 points) For which of the variables is it of scientific interest to assess skewness? Do any of those variables appear to be skewed? Briefly explain your reasons.

Ans: Skewness is of interest for the uncensored continuous variables: age, bmi, packyrs, physact, ldl, sbp.

Both packyrs and physact appear markedly skewed to the right: the mean is very different from the median, the maximum is much further from the median than is the minimum, the 75th percentile is further from the median than is the 25th percentile, and the standard deviation is extremely large relative to the mean for these nonnegative random variables.

Age, bmi, and sbp appear to have some large values (the maximum is further from the median than is the minimum), but they do not appear to be marked outliers, because there are no other signs of extreme skewness.

4. (25 points total) The following graph displays a scatterplot of serum creatinine (CRT: a measure of kidney function) versus forced expiratory volume (FEV: a measure of lung function) from the brain imaging study. Superimposed on this scatterplot is a lowess curve.



a. (5 points) Briefly summarize the observations you would make from this graph.

Ans: There appear to be three to four creatinine measurements that are much more extreme (outliers) than that for others having the same FEV.

There seems to be a very slight trend toward higher average CRT for subjects with higher FEV.

That trend in average CRT appears to be well approximated by a straight line.

Apart from the outlying measurements, there does not appear to be heteroscedasticity, though this is very difficult to judge due to the many observations in the middle of the plot.

b. (5 points) The correlation between creatinine and FEV was computed to be 0.223 (95% confidence interval 0.152 to 0.291). What does this confidence interval suggest about an association between creatinine and FEV in the population? (Hint: if creatinine and FEV were independent, what would be the correlation between them?)

Ans: The 95% CI for the correlation does not include 0, hence we can reject the null hypothesis of no linear association between CRT and FEV in this population, in favor of there being some linear trend. *(Because we unfortunately only have the correlation, it is difficult to judge the clinical importance of this trend. I wish I had the slope estimate more precisely quantified than I can do from the graph.)*

c. (5 points) When analyses were done separately for men and women, the correlation between creatinine and FEV was -0.006 for women and 0.001 for men. How might you explain the difference between the results in the combined sample and the results observed in the individual sexes? (Speculate on what relationships there might be between sex, creatinine, and FEV that would explain these results.)

Ans: Within sex groups, there does not appear to be a linear association between CRT and FEV. This would suggest that one sex might tend toward lower CRT and lower FEV, and the other sex might tend toward higher CRT and higher FEV. Knowing what we know about FEV and height and sex and height, it would seem reasonable to guess that men average higher CRT and higher FEV than women, but within each sex, CRT and FEV are relatively independent of each other. This could be because men are more prone to kidney disease, but women are more prone to lung disease. Alternatives include the relationship between body size and FEV and the association between more muscle mass and with higher creatinine levels.

d. (5 points) What do the analyses presented in part c suggest about the role of sex (e.g., effect modification, confounding, precision, …) when our interest is in investigating an association between creatinine and FEV?

Ans: Sex appears to be a confounder. *It appears to be associated with CRT and associated with FEV and not in a causal pathway of interest. The two sexes appear to have quite similar associations between CRT and FEV (i.e., no association), and thus it is not an effect modifier.*

e. (5 points) What do the analyses presented in part c suggest about the role FEV might play (e.g., effect modification, confounding, precision, …) when our interest is in investigating an association between creatinine and sex?

Ans: FEV would not be an effect modifier, confounder, or precision variable. It does not appear to be associated with CRT after adjusting for sex.

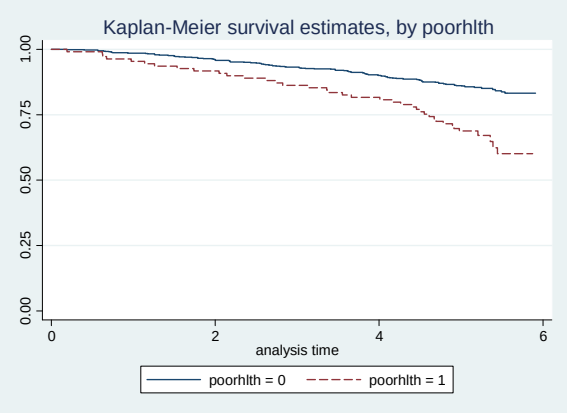
5. (10 points) The following table presents descriptive statistics for selected variables according to whether the patient was in generally poor health (*genhlth* = 4 or 5) or in generally good (*genhlth* = 1, 2, or 3).

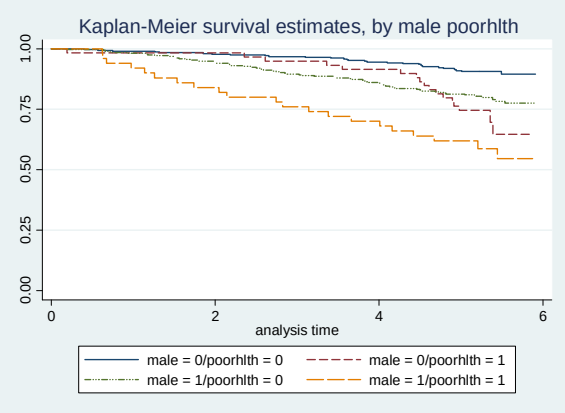
Variable	N	Mean	SD	Min	25th %ile	Median	75th %ile	Max
Patients in Generally Good Health								
age	626	74.4	5.3	65	70	73	78	99
sex	626	0.505	0.500	0	0	1	1	1
obstime	626	4.99	1.00	0.19	5.05	5.15	5.61	5.91
status	626	0.152	0.359	0	0	0	0	1
Patients in Generally Poor Health								
age	109	75.7	6.0	66	71	74	79	91
sex	109	0.459	0.501	0	0	0	1	1
obstime	109	4.63	1.37	0.19	4.61	5.08	5.39	5.88
status	109	0.349	0.479	0	0	0	1	1

a. (10 points) How would you use the above descriptive statistics to assess whether the presence of generally poor health is associated with shortened survival?

Ans: I would not. We have censored survival times, and thus we have to look at Kaplan-Meier estimates. *(If I knew that the censoring distribution were similar for all health groups, maybe I could interpret this data a little, because then I would know that the observation of death was not confounded by the observation time. But not knowing this, I am leery of drawing any conclusions.)*

6. (30 points) The following are results from a Kaplan-Meier analyses of the time to death within strata defined by the presence of generally poor health and sex.





	Generally Good Health			Generally Poor Health		
		Survival Probability			Survival Probability	
	n	3 Year	5 Year	n	3 Year	5 Year
All Patients	626	0.931	0.861	109	0.862	0.688
Female	310	0.968	0.910	59	0.949	0.746
Male	316	0.896	0.813	50	0.760	0.620

a. (10 points) Based on the above statistics, would you conclude that there is overall an association between the presence of generally poor health and the probability of survival? Provide statistics to quantify your answer.

Ans: Graphically, the survival probability curve is lower for patients in poor health. The probability of surviving is an absolute difference of 6.9% lower for those in poor health at 3 years and 17.3% lower at 5 years.

b. (10 points) Based on the above statistics, would you conclude that the sex modifies any association between the presence of generally poor health and survival? Provide statistics to quantify your answer.

Ans: When considering the sexes separately:

For females, the probability of surviving is an absolute difference of 1.9% lower for those in poor health at 3 years and 16.4% lower at 5 years.

For males, the probability of surviving is an absolute difference of 13.6% lower for those in poor health at 3 years and 19.3% lower at 5 years.

Certainly the striking difference for the sexes in the association between general health status and 3 year survival probability would be suggestive of effect modification over that time frame. If 5 year survival were the primary interest, there is less striking effect modification.

We could also have looked at the sex effect within categories defined by health status:

For subjects in generally good health, the probability of suriviving is an absolute difference of 7.2% lower for males at 3 years and 9.7% lower at 5 years.

For subjects in generally poor health, the probability of surviving is an absolute difference of 18.9% lower for males at 3 years and 12.6% lower at 5 years.

Again, the evidence for effect modification is more striking for 3 year survival probailities.

It should be noted that I used wording to disambiguate what I meant by being some percentage lower:

If one group had survival probability of 80%, and the other had survival probability of 60%, that is an "absolute" (additive) difference of 20% lower, but a "relative" (multiplicative) difference of 25% lower (0.2 / 0.8 = 0.25) I find it best to explicitly mention which I mean. Sometimes, I even give both scales to make it clear..

c. (10 points) Based on the above statistics, would you conclude that sex confounds any association between the presence of generally poor health and survival? Provide statistics to quantify your answer.

Ans: Among females, 59 / 369 = 16% were in generally poor health. Among males, 50 / 366 = 13.7% were in generally poor health. I do not consider the differences in these percentages evidence of an association between sex and general state of health in the sample, and hence there is not confounding. *This is true despite the evidence of a strong association between sex and survival (men are so fragile).*

7. (30 points total) Suppose we are interested in studying whether C reactive protein can accurately predict the presence of severe narrowing of the coronary arteries (blood supply to the heart muscle). The "gold standard" for the diagnosis of severe narrowing of the coronary arteries would be based on a more invasive radiologic examination (coronary angiography).
   - Suppose we sample 500 subjects with radiographic evidence of severe narrowing of the arteries. Among these subjects we find that 375 subjects have a C reactive protein level greater than 3 mg/dl.
   - Suppose we also sample 1000 subjects whose radiographic studies show no evidence of severe narrowing of the arteries. Among these subjects we find that 150 subjects have a C reactive protein level greater than 3 mg/dl.

   a. (5 points) Can the above data be used to estimate the probability of severe narrowing of the coronary arteries in the population? If so, provide the estimate. If not, briefly explain why not.

Ans: No. The study design fixed the relative size of the healthy (no severe narrowing) and diseased (severe narrowing) populations.

   b. (5 points) Can the above data be used to estimate the sensitivity of a high C reactive protein in predicting severe narrowing of the coronary arteries? If so, provide the estimate. If not, briefly explain why not.

Ans: Yes, the probability of a positive test (CRP > 3 mg/dl) among the diseased (severe narrowing) is 375 / 500 = 0.75.

   c. (5 points) Can the above data be used to estimate the specificity of a high C reactive protein in predicting severe narrowing of the coronary arteries? If so, provide the estimate. If not, briefly explain why not.

Ans: Yes the probability of a negative test (CRP < 3 mg/dl) among the healthy (no severe narrowing) is (1000 – 150) / 1000 = 0.85.

d. (5 points) Can the above data be used to estimate the positive predictive value of a high C reactive protein in predicting severe narrowing of the coronary arteries? If so, provide the estimate. If not, briefly explain why not.

Ans: No. The study design fixed the relative size of the healthy (no severe narrowing) and diseased (severe narrowing) populations.

e. (5 points) Can the above data be used to estimate the negative predictive value of a high C reactive protein in predicting severe narrowing of the coronary arteries? If so, provide the estimate. If not, briefly explain why not.

Ans: No. The study design fixed the relative size of the healthy (no severe narrowing) and diseased (severe narrowing) populations.

f. (5 points) Suppose the probability of severe narrowing of the coronary arteries is 20% in the population. What is the positive predictive value of high C reactive protein? Explain how you derived your answer.

Ans: We use Bayes' Rule:

$$
\begin{aligned}
PPV &= \Pr(Disease \mid Positivity) \\
&= \frac{\Pr(Positivity \mid Disease)\Pr(Disease)}{\Pr(Positivity \mid Disease)\Pr(Disease) + \Pr(Positivity \mid Healthy)\Pr(Healthy)} \\
&= \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - specificity) \times (1 - prevalence)} \\
&= \frac{0.75 \times 0.20}{0.75 \times 0.20 + 0.15 \times 0.80} = 0.556
\end{aligned}
$$

Grades:

Maimum Possible: **146**

Highest Achieved: **143**

Mean (SD) **111 (20.2)**

Percentile	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Grade	**83**	**93**	**101**	**107**	**114**	**121**	**124**	**131**	**134**